Mineração de Dados em Biologia Molecular

Preparação de dados

André C. P. L. F. de Carvalho
Monitor: Valéria Carvalho

Tópicos

- Dados
- Caracterização de dados
  - Instâncias e Atributos
  - Tipos de Dados
- Exploração de dados
  - Dados univariados
  - Medidas de localidade, espalhamento e distribuição
  - Dados multivariados
  - Visualização

Conjuntos de dados

Atributos de entrada (preditivos)

Nome	Temp.	Idade	Peso	Altura	
João	37	70	94	190	Saudável
Maria	38	65	60	172	Doente
José	39	19	70	185	Doente
Sílvia	38	25	65	160	Saudável
Pedro	37	70	90	168	Doente

Tipos de atributos

- Nominal
  - Ex.: cor, código de identificação, profissão
- Ordinal
  - Ex.: gosto (ruim, médio, bom), dias da semana
- Intervalar
  - Ex.: data, temperatura em Celsius
- Racional
  - Ex.: peso, tamanho, idade

Exemplo

Nome	Temp	Enjôo	Mancha	Dor	Salário	Diagnóstico
João	37.7	sim	pequena	sim	1000	doente
Pedro	37.0	não	pequena	não	1100	saudável
Maria	38.2	sim	grande	não	600	saudável
José	39.0	não	pequena	sim	2000	doente
Ana	37.3	não	grande	sim	1800	saudável
Leila	37.7	não	grande	sim	900	doente

Tipos de atributos

- Nominal (=, ≠)
  - Valores são apenas nomes diferentes
- Ordinal (<, >)
  - Existe uma relação de ordem entre valores
- Intervalar (+, -)
  - Diferença entre valores faz sentido
- Racional (*, /)
  - Razão e diferença entre valores fazem sentido

Exercício

- Definir o tipo dos seguintes atributos:
  - Renda mensal
  - Número de palavras de um texto
  - Fotografia
  - Número de RG
  - Data de nascimento
  - Código de disciplina
  - Posição em uma corrida

Quantidade de valores

- Atributos também se distinguem pela quantidade de valores
  - Discretos
    - Número finito ou infinito e enumerável de valores
      - Ex.: código postal, quantidade de algum elemento
    - Caso especial: valores binários
  - Contínuos
    - Assumem valores contínuos, como números reais
      - Ex.: temperatura, peso, distância

Exploração de dados

- Exploração preliminar dos dados facilita entendimento de suas características
- Principais motivações:
  - Ajudar a selecionar a melhor técnica para pré-processamento ou modelagem
- Estatística descritiva
- Visualização

Estatística descritiva

- Descreve dados
- Produz valores que resumem características de um conjunto de dados
  - Na maioria das vezes por meio de cálculos simples

Estatística descritiva

- Pode capturar:
  - Frequência
  - Localização ou tendência central
    - Ex.: Média
  - Dispersão ou espalhamento
    - Ex.: Desvio padrão
  - Distribuição ou formato

Frequência

- Proporção de vezes que um atributo assume um dado valor
  - Para um determinado conjunto de dados
  - Muita usada para dados categóricos
  - Ex.: Em um conjunto de dados médicos, 40% dos pacientes moram no interior

Exemplo
Febre Idade Mancha Dor Diagnóstico
sim 23 grande sim doente
não 9 pequena não saudável
sim 61 grande não saudável
sim 32 pequena sim doente
sim 21 grande sim saudável
não 48 grande sim doente
66% das manchas são manchas grandes
André C P L F de Carvalho 13

Medidas de localidade
Dados simbólicos ou categóricos
Moda
Dados numéricos
Média
Mediana
Percentil
André C P L F de Carvalho 14

Exemplo
Febre Idade Mancha Dor Diagnóstico
sim 23 grande sim doente
não 9 pequena não saudável
sim 61 grande não saudável
sim 32 pequena sim doente
sim 21 grande sim saudável
não 48 grande sim doente
Moda para o atributo mancha: grande
André C P L F de Carvalho 15

Média
Pode ser calculada facilmente
$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Problema: sensível a outliers
André C P L F de Carvalho 16

Mediana
Menos sensível a outliers que média
Necessário ordenar valores
$mediana(x) = \begin{cases} x_{(r+1)} \text{ se n é ímpar } (n = 2r+1) \\ \frac{1}{2}(x_r + x_{(r+1)}) \text{ se n é par } (n = 2r) \end{cases}$
André C P L F de Carvalho 17

Média versus Mediana
Média é um bom indicador do meio de um conjunto de valores quando os valores estão distribuídos simetricamente
Mediana indica melhor o meio
Se distribuição é oblíqua (assimétrica)
Skewed
Se existem outliers
André C P L F de Carvalho 18

Média Podada

- *Trimmed mean*
- Minimiza problema da média descartando exemplos nos extremos
  - Define porcentagem p dos exemplos a serem eliminados
  - Ordena os dados
  - Elimina (p/2)% dos exemplos em cada extremidade

Exercício

- Dado o conjunto de dados {1, 2, 3, 4, 5, 80}, calcular:
  - Média
  - Mediana
  - Média podada com p = 33%

Quartis e Percentis

- Mediana divide os dados ao meio
- Outras medidas usam pontos de divisão diferentes
  - Quartis dividem um conjunto ordenado de dados em quartos
    - Primeiro quartil, Q_1, é o valor da observação para a qual 25% do conjunto (amostra) tem valor menor ou igual a ela
      - Também é o valor da amostra 25º percentil
    - Segundo quartil, Q_2, = mediana

Percentis

- Valor da amostra 100pº percentil é uma valor em que:
  - Pelo menos 100p% das observações possuem um valor menor ou igual a ela
  - Pelo menos 100(1-p)% das observações tem um valor igual ou acima
- Mediana é o 50º percentil
  - Para cálculo, usar fórmula da mediana

Cálculo dos Percentis

- Ordenar os valores
  - Posição do p-percentil:

$$posição = \left\lceil \frac{p}{100} \times n + \frac{1}{2} \right\rceil$$

    - Arredonda posição para o valor inteiro mais próximo
    - Retornar o valor nessa posição

Exemplo

- Obter os quartis e a 95º percentil para o conjunto de dados abaixo:

6.2	7.67	8.3	9.0	9.4
9.8	10.5	10.7	11.0	12.3

Exemplo

- Obter os quartis e a 95º percentil para o conjunto de dados abaixo:

6.2	7.67	8.3	9.0	9.4
9.8	10.5	10.7	11.0	12.3

Q_1: np = 0.25x10+ 0.5= 3
usar o terceiro valor: Q_1 = 8.3
Q_2: np = 0.5x10 + 0.5 = 5.5
para a mediana, usar a média entre o quinto e o sexto valor: Q_2 = 9.6
Q_3: np = 0.75x10 + 0.5= 8
usar o oitavo valor: Q_3 = 10.7
$P_{0.95}$: np = 0.95x10 + 0.5= 10
usar o décimo valor: $P_{0.95}$ = 12.3

Exercício

- Calcular quartis inferior e superior e o 60º percentil para os valores
  - 16, 25, 4, 18, 11, 13, 20, 8, 11 e 9

Exercício

- Calcular quartis inferior e superior para os valores
  - 16, 25, 4, 18, 11, 13, 20, 8, 11 e 9
  - 4, 8, 9, 11, 11, 13, 16, 18, 20, 25
  - Q_1 =
  - Q_3 =
  - 60º percentil =

Exercício

- Dados os números abaixo, calcular a mediana, o primeiro quartil e o segundo quartil
  - 23, 7, 12, 6, 10
  - 23, 7, 12, 6, 10, 7

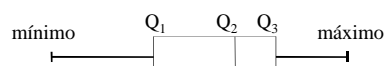
Exercício

- Obter os quartis, o 30º percentil e o 95º percentil para o conjunto de dados:

3,20	11,70	13,64	15,60	15,89	28,44	29,07
37,34	41,81	43,35	43,94	49,51	49,82	51,20
51,43	52,47	53,72	53,92	54,03	56,89	63,80
66,40	68,64	70,15	70,98	74,52	76,68	77,84
80,91	84,04	85,70	86,48	88,92	89,28	91,36
91,62	98,79	102,39	104,21	124,27		

Boxplot

- Gráfico que resume informações dos quartis

mínimo Q_1 Q_2 Q_3 máximo

Medidas de Espalhamento

- Medem dispersão ou espalhamento de um conjunto de valores
- Indicam se os dados estão:
  - Amplamente espalhados ou
  - Relativamente concentrados em torno de um ponto (ex. média)
- Medidas comuns
  - Intervalo
  - Variância
  - Desvio padrão

Intervalo

- Medida mais simples, mostra espalhamento máximo
- Sejam $\{x_1, \ldots, x_n\}$ os valores do atributo x para n objetos

$$r(x) = \max(x) - \min(x)$$

- Pode não ser uma boa medida
  - Se a maioria dos valores forem próximos de um ponto, com um pequeno número de valores extremos

Variância

- Medida preferida para analisar espalhamento dos dados

$$\text{var}(v) = \frac{1}{n-1}\sum_{i=1}^{n}(v_i - \bar{v})^2$$

  - Denominador *m-1*: correção de Bessel, usada para uma melhor estimativa da variância verdadeira
  - Desvio padrão: raiz quadrada da variância

Momento

- Estima parâmetros de uma população de valores

$$mom_k = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^k}{(n-1)} \quad \text{ou} \quad \mu_k = \sum_{i=1}^{n}(x_i - \mu)^k \, p(x_i)$$

$$p(x_i) = f_i$$

  - Valor de k define a medida de momento

Momento

- K-ésimo momento central ou centrado
  - K=1: 0 (primeiro momento em torno da origem – primeiro momento central)
  - K=2: variância (segundo momento central)
  - K=3: obliquidade (terceiro momento central)
  - K=4: curtose (quarto momento central)

Obliquidade

- Terceiro momento (*Skewness*)
  - Mede a simetria da distribuição dos dados em torno da média
    - Distribuição simétrica tem a mesma aparência à direita e à esquerda do ponto central

$$Obl = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)\sigma^3}$$

Dividido por σ^3 para tornar a medida independente de escala

$$\mu_3 = \frac{1}{\sigma_3}\sum_{i=1}^{n}(x_i - \mu)^3 \, p(x_i)$$

Curtose

- Quarto momento (*Kurtosis*)
  - Medida de dispersão que captura o achatamento da função de distribuição
    - Verifica se os dados apresentam um pico ou são achatados em relação a uma distribuição normal

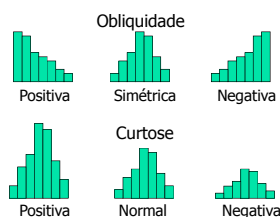
$$Curt = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)\sigma^4}$$

Curtose

- Para uma distribuição normal padrão, *Curt = 3*
  - Média = 0 e desvio padrão = 1
- Para que a distribuição normal padrão tenha curtose = 0, usa-se

$$Curt = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)\sigma^4} - 3$$

Histograma

- Melhor forma para verificar graficamente curtose e obliquidade

Exercício

- Obter o valor dos 4 primeiros momentos centrais para os dados:

3,20 11,70 13,64 15,60 15,89 28,44 29,07

Dados Multivariados

- Aqueles que possuem vários atributos
- Medidas de localização
  - Podem ser obtidas calculando medida de localização de cada atributo separadamente
    - Ex.: média, mediana, ...
  - Média dos objetos de um conjunto de dados com *m* atributos é dada por:

$$\bar{x} = (\bar{x}_1, \ldots, \bar{x}_m)$$

Dados Multivariados

- Medidas de espalhamento
  - Podem ser calculadas para cada atributo independentemente dos demais
    - Usando qualquer medida de espalhamento
  - Variáveis contínuas
    - Espalhamento de um conjunto de dados é melhor capturado por uma matriz de covariância
      - Cada elemento é a covariância entre dois atributos

Dados Multivariados

- Matriz de covariância *S* para um conjunto de dados com *n* objetos

$$s_{ij} = \text{covariância}(x_i, x_j)$$

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

Onde:
$\bar{x}_i$: Valor médio do i-ésimo atributo
x_{ki}: Valor do i-ésimo atributo para o k-ésimo objeto

- Obs: covariância (x_i, x_i) = variância (x_i)
  - Matriz de covariância tem em sua diagonal as variâncias dos atributos

Exercício

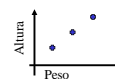
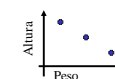
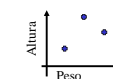
- Calcular a matriz de covariância para o conjunto de dados:

Peso	Altura	Temperatura
73,2	170	37,5
67,5	165	38
90	190	37,2
49	152	37,8

Dados Multivariados

- Covariância de dois atributos
  - Mede o grau com que os atributos variam juntos
    - Depende da magnitude dos atributos
    - Valor próximo de 0:
      - Atributos não têm um relacionamento linear
    - Valor positivo:
      - Atributos diretamente relacionados
        - Quando o valor de um atributo aumenta, o do outro também aumenta

Exemplo

Peso	Altura
60	170
70	180
80	190

Peso	Altura
60	190
70	180
80	170

Peso	Altura
60	170
70	190
80	180

Dados Multivariados

- Covariância de dois atributos
  - Não é possível avaliar o relacionamento entre dois atributos olhando apenas a covariância
    - Correlação entre dois atributos dá uma indicação mais clara da força da relação linear entre eles
      - Mais popular que covariância

Dados Multivariados

- Correlação
  - Indica força da relação entre dois atributos
  - Matriz de correlação R

$$r_{ij} = \text{correlação}(x_i, x_j) = \frac{\text{covariância}(x_i, x_j)}{s_i s_j}$$

Onde:
x_i: i-ésimo atributo
s_i: Variância do atributo x_i

  - Obs: correlação (x_i, x_i) = 1
    - Elementos da diagonal tem valor 1
    - Demais elementos têm valor entre –1 e +1

Exercício

- Calcular a matriz de correlação para o conjunto de dados:

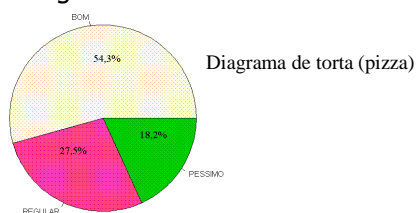
Peso	Altura	Temperatura
73,2	170	37,5
67,5	165	38
90	190	37,2
49	152	37,8

Outras formas de sumarizar dados

- Visualização gráfica
  - Em vários casos, facilita compreensão de aspectos mais complicados dos dados
  - Ex.: Histogramas

Diagrama de torta

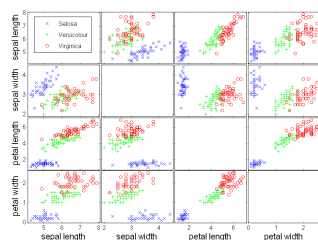
- Frequências relativas podem ser vistas no diagrama circular

Diagrama de torta (pizza)

Scatter Plot

- Usado para ilustrar correlação linear
- Cada objeto é associado a uma posição em um gráfico
  - Valores dos atributos definem sua posição
  - Os valores podem ser inteiros ou reais
- Matrizes de scatter plot resumem relação entre vários pares de atributos

Scatter Plot

- Matriz para atributos do conjunto iris

Diferentes classes são indicadas por cores diferentes

Faces de Chernoff

- Criado por Herman Chernoff
- Mapeia os valores dos atributos para imagens mais familiares: faces
  - Cada objeto é representado por uma face
  - Cada atributo é associado a uma característica específica de uma face
- Baseia-se na habilidade humana de distinguir faces

Febre	Idade	Mancha	Dor	Diagnóstico
sim	23	grande	sim	doente
não	9	pequena	não	saudável
sim	61	grande	não	saudável
sim	32	pequena	sim	doente
sim	21	grande	sim	saudável
não	48	grande	sim	doente